

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

داده کاوی

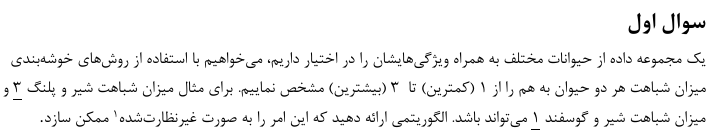
(بهار ۱۴۰۱)

تمرین دوم

**محمد چوپان ۹۸۳۱۱۲۵**

**بخش تئوری :**

**سوال اول :**

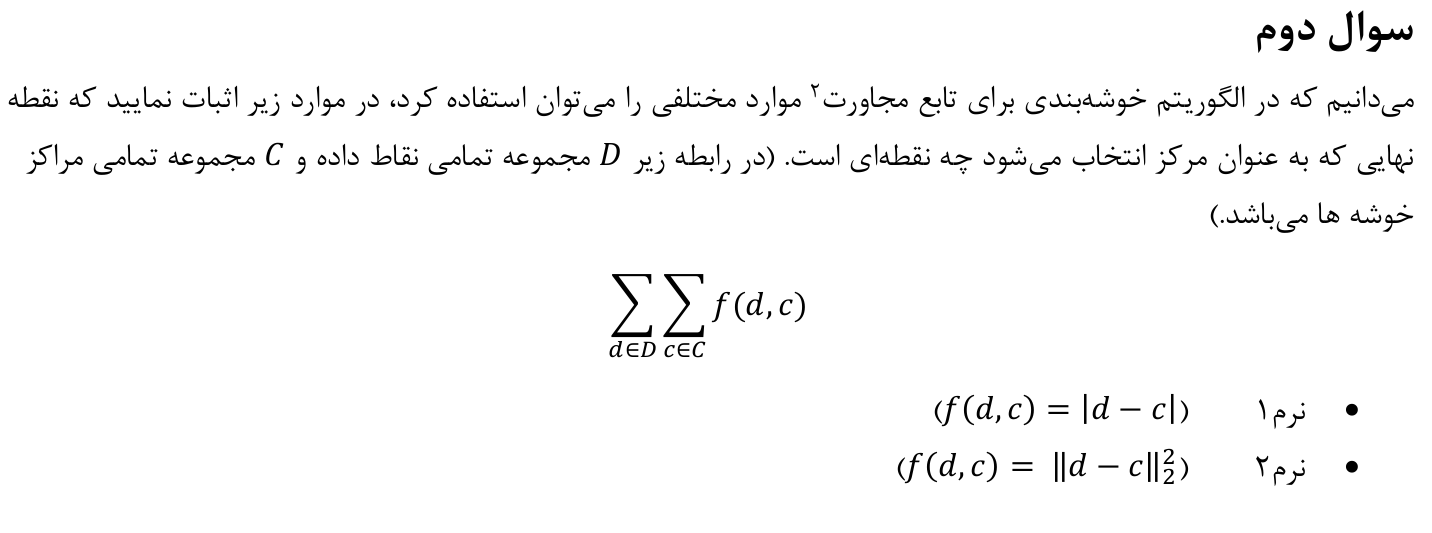
****

**پاسخ :**

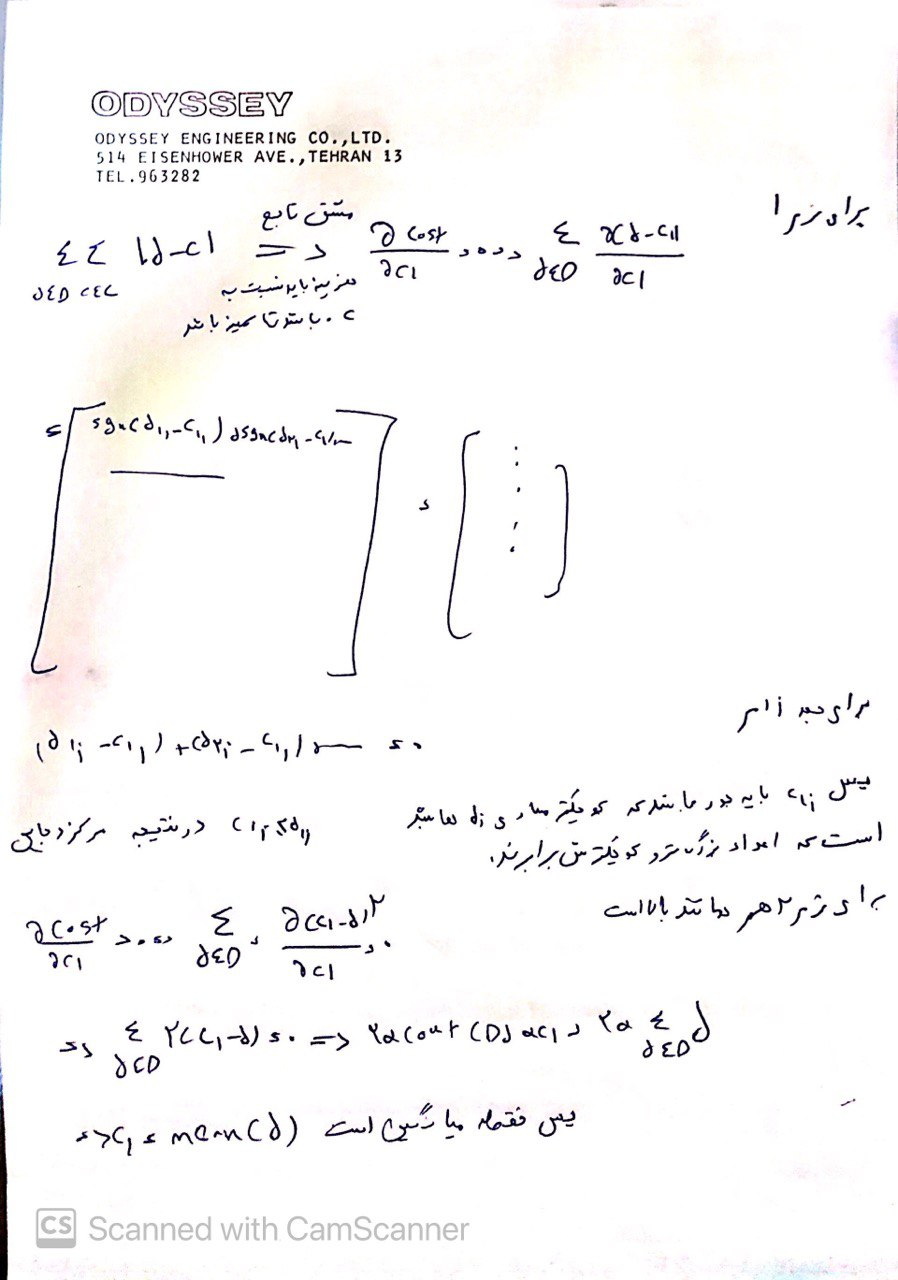
**‌**برای انجام این کار ابتدا میتوانیم داده‌هارا پیش پردازش کنیم و داده ها را به صورت ماتریسی در بیاوریم.

سپس برای انتخاب ویژگی میتوانیم داده های غیر عددی را به داده های عددی در بیاوریم و ویژگی های مهم آن را انتخاب کنیم در صورت نیاز میتوانیم از الگوریتم کاهش بعد غیر نظارتی PCA نیز استفاده کنیم.در نهایت میتوانیم با استفاده از الگوریتم خوشه بندی kmeans داده های خود را دسته بندی کنیم .می توانیم k را در اینجا ۳ در نظر بگیریم . حال شباهت بین داده های یک خوشه را ۳ در نظر میگیریم برای به دست آوردن شباهن بیت داده های ۲ خوشه مختلف نیز میتوانیم معیار فاصله را در نظر بگیریم و یک عدد بین ۱ تا ۲ به آن اختصاص دهیم .یک فرایند دیگر نیز این است که ابتدا ماتریس ویژگی ها را تهیه کنیم و پس از مراحل پیش پردازش مانند استاندارد سازی از الگوریتم های کاهش بعد مانند PCA استفاده کنیم و در نهایت شباهت بردار های ویژگی را با استفاده از معیاری مانند فاصله اقلیدسی حساب کنیم و داده ها را بین ۱ تا ۳ نرمال سازی کنیم.

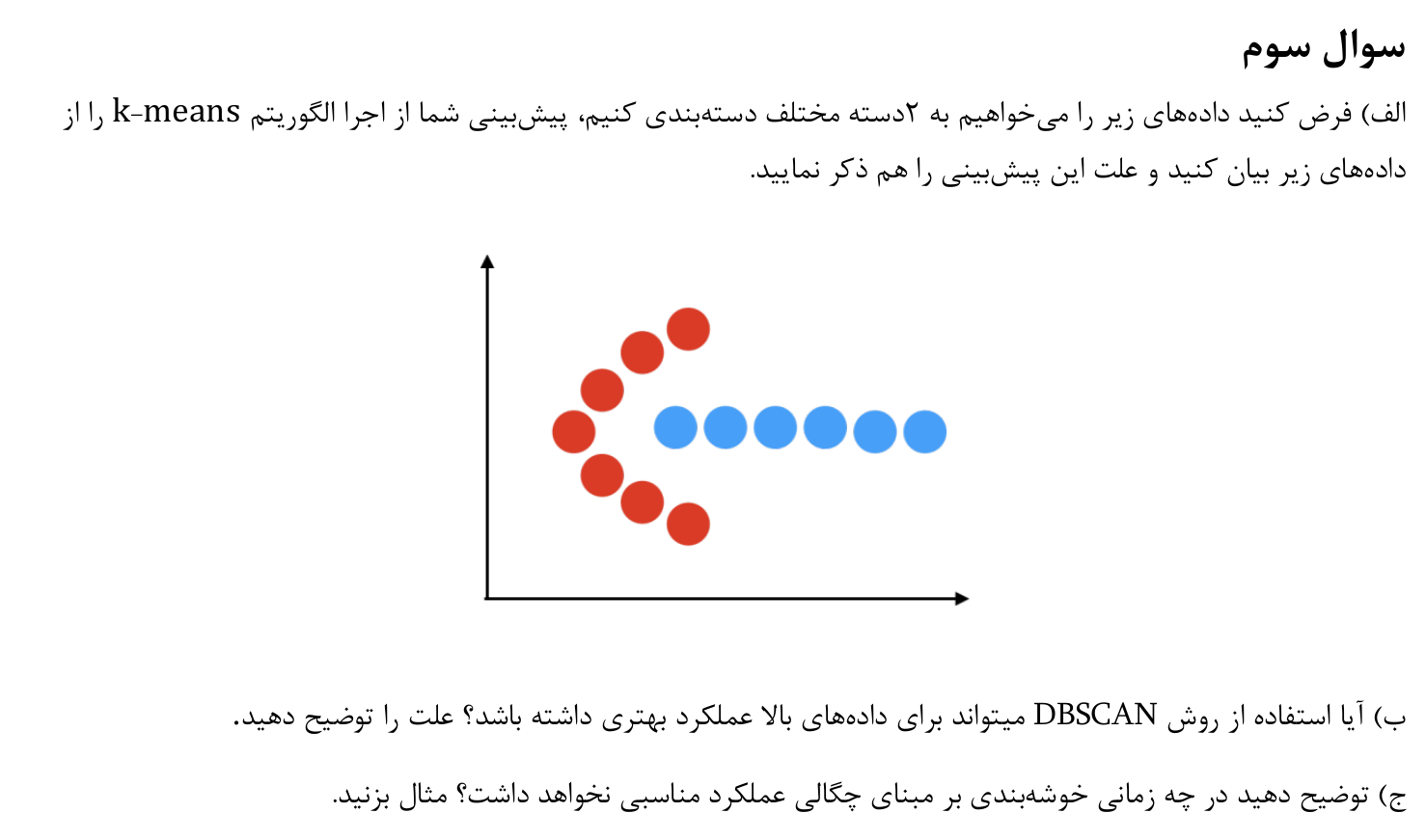
**سوال دوم :**

****

**پاسخ :**

****

**سوال سوم:**

****

**پاسخ :**

پیش بینی ما مانند شکل زیر است به دلیل اینکه سعی میکند بیشترین نقاط در کمترین فاصله را داشته باشد که عملکرد خوب و درستی هم نیست. پیاده سازی انجام شده نیز پیش بینی ما را تایید میکند. همچنین اگر مراکز ابتدایی هر رندومی باشند به وسط دو دسته میل پیدا میکند و حرف ما را تایید میکند.

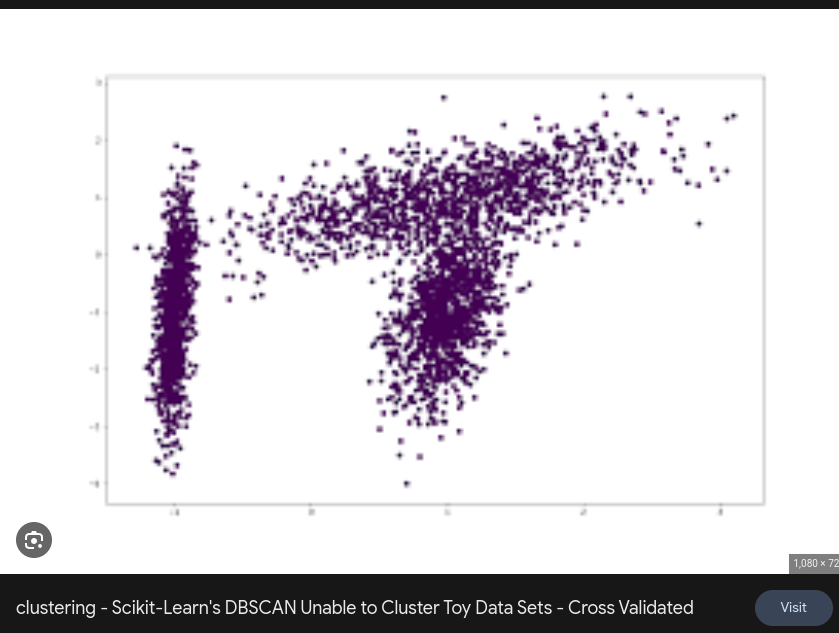
که دو نقطه کوچک شکل مراکز نهایی ما هستند.

****

**ب:** بله زیرا روش DBSCAN یک روش مبتنی بر چگالی می باشد و اگر EPS را مقدار معقولی قرار دهیم یعنی بیشتر از حداکثر فاصله آبی ها با هم و قرمز ها باهم و کمتر از حداقل فاصله آبی و قرمز با یکدیگر می تواند این دو دسته را به خوبی جدا کند .این الگوریتم به دلیل اینکه با فاصله مراکز کار نمیکند و از یک مقدار heuristic استفاده میکند در شکل هایی که این حالت را دارند مانند دو دایره تو در تو عملکرد بهتری دارد.

**ج:**

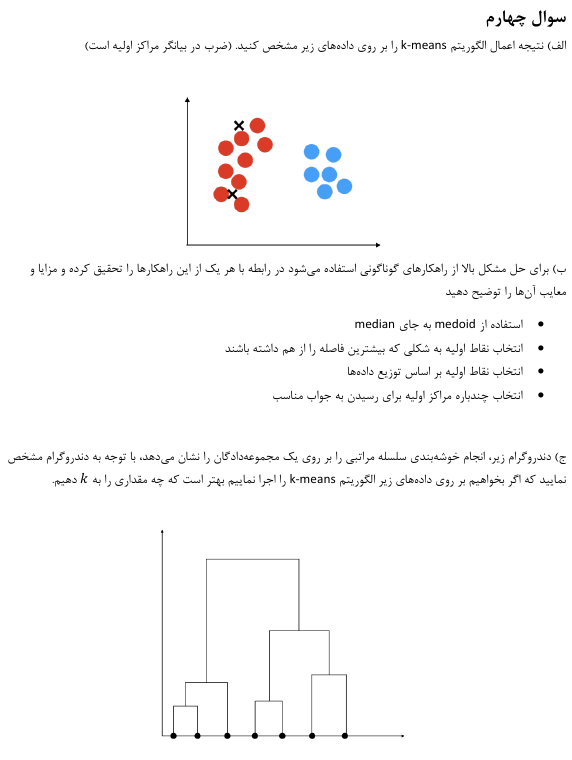
زمانی که چگالی داده های ما در نقاط متفاوتی درون یک خوشه متفاوت باشد یعنی درون یک خوشه در یکجا خیلی زیاد و در یکجا خیلی کم باشد و به اصلاح توضیع متوازنی نداشته باشد. مانند شکل زیر یا زمانی که داده های ما خیلی یعد های متفاوتی داشته باشد نمیتواند به خوبی عمل کند.

****

به دلیل مقدار زیاد noise و outlier علی رغم اینکه DBSCAN در نویز ها خوب عمل میکند اما اینجا موفق نشده است .

در کل الگوریتم های مبتنی بر چگالی این مشکل را دارند.

**سوال چهارم :**

****

**پاسخ :**

**الف:**

نتایج برای الگوریتم K =۲ برابر است با شکل زیر دلیل آن هم این است که مرکز پایینی وسط ۴ تا آبی و ۴ تا قرمز است و مرکز بالایی نیز به دلیل بیشتر بودن قرمز ها به سمت قزمر ها مایل شده است. علت این اتفاق نیز نزدیک بود مراکز اولی و درست انتخاب نشدن آن ها است.

****

**ب:**

**استفاده از medoid به جای median :**

مزایا : انتخاب مرکز از medoid ها از این نظر بهتر است که داده های خارجی یا همان outlier ها را بهتر تشخیص داده و نسبت به آن ها کمتر حساس است. و اینکه مدوید به دلیل اینکه از خود داده ها است برای داده های غیر عددی نیز مناسب تر است. همچنین استفاده از مدوید باعث پایداری الگوریتم می‌شود و تاقیرات نویز ها را نیز کمتر می‌کند.

معایب: محاسبات آن بیشتر است و از نظر محاسبات هزینه زیاد بر ما تحمیل می شود و بعضی وقتی تشخیص خود مدویدد نیز سخت است.

**انتخاب نقطه اولیه به شکلی که بیشترین فاصله را داشته باشند:**

مزایا : برای داده هایی که خیلی با هم فاصله دارند مناسب تر است و همچنین در بهینه های محلی گیر نمیکند لزوما

معایب : به داده های بیرونی و نویز ها بسیار حساس تر می شود و همچنین باعث میشود که تعداد تکراری بیشتری داشته باشیم. و اگر خوشه های ما با هم مقداری همپوشانی داشته باشند نتایج جالبی نخواهد داشت.

**انتخاب نقاط اولیه بر اساس توزیع داده ها :**

مزایا: احتمال اینکه در مینیمم محلی گیر کنیم کمتر است و معمولا برای داده های است که یک توزیع یکسانی ندارند

معایب : محاسبات بسیار زیادی دارد و مجدد بر نویز ها نیز حساس تر می شود و همچنین این حالت وجود دارد که به علت اینکه داده ها توزیع یکسانی ندارند در خوشه های خاصی تمرکز بیشتری داشته باشد.

**انتخاب چند باره مراکز اولیه برای رسیدن به جواب مناسب :**

مزایا: یکی از بهترین های روش ها این است به دلیل اینکه هم خطا را کمتر میکند هم نقش نقاط بیرونی را کمرنگ تر میکند و اینکه تقریبا احتمال گیر کردن در مینیموم محلی وجود ندارد

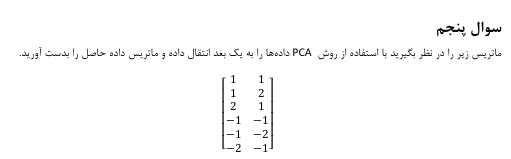
معایب: تعداد محاسبات ما خیلی بالا می رود و یک روش مشخص برای اتمام کار وجود ندارد لزوما

**د :**

****

بهتر است که یکی از این دو عدد را انتخاب کنیم هم بر اساس روش elbow هم اینکه ارتفاع نمودار دندوگرام میزان شباهت داده ها را نشان میدهد و اینجا هر چی بیشتر باشد داده ها متفاوت ترند برای دسته بندی این اعداد نتایج بهتری میدهند. بین ۳ یا ۴ یک داده تنها تفاوت دارد که میتوان ۳ را برای محاسبات کمتر نیز انتخاب کرد.

**سوال پنجم :**

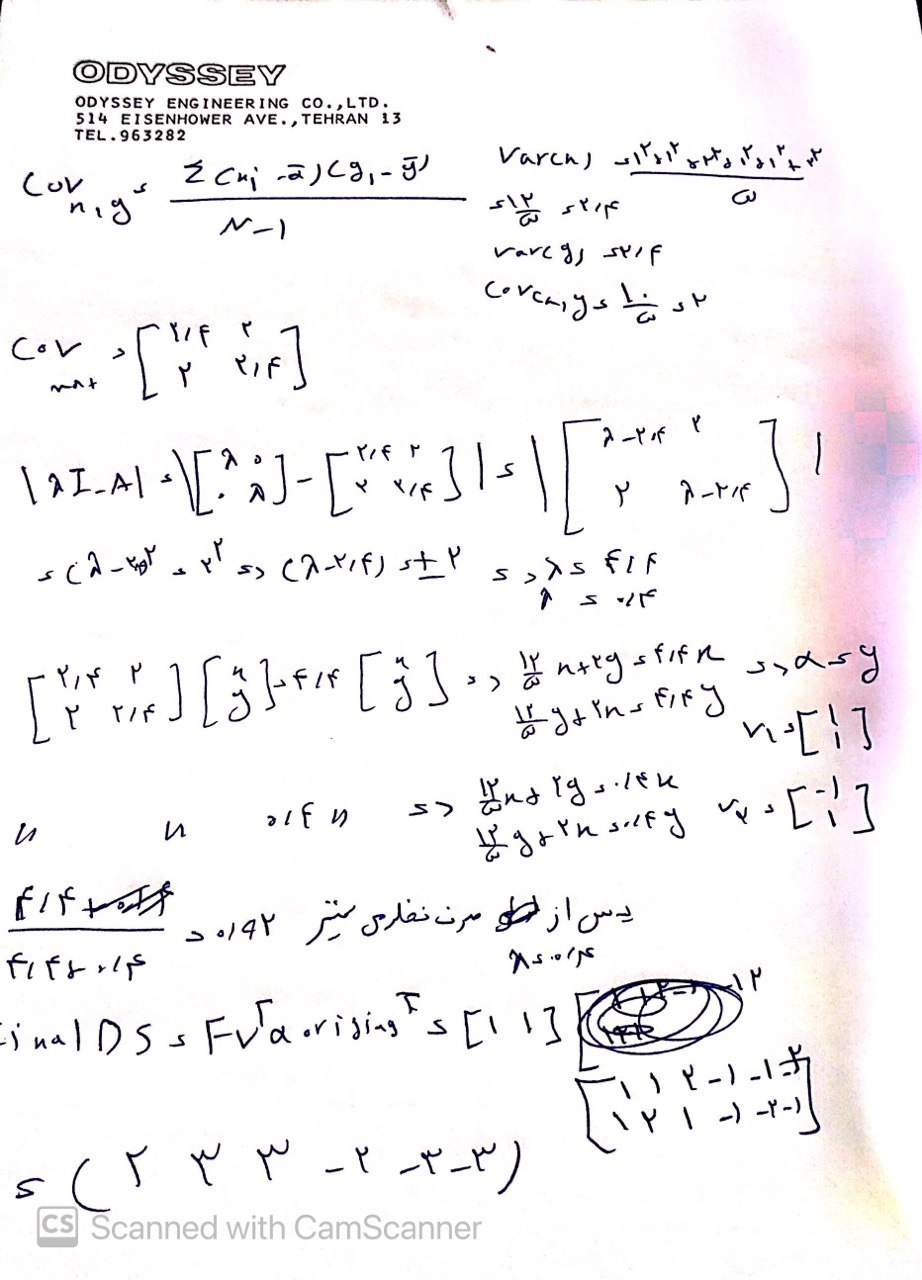


**پاسخ :**

ابتد در مرحله اول باید داده ها را استاندارد کنیم به دلیل اینکه میانگین ما برابر ۰ است نیازی به انجام این مرحله نداریم.

در مرحله بعدی باید ماتریکس کواریانس را محاسبه کنیم که بقیه محاسبات در تصویر زیر وجود دارد.

در پاسخ نهایی تفاوتی ندارد اما مخرج cov ابتدا n بود سپس به n-1 تغییر دادم.

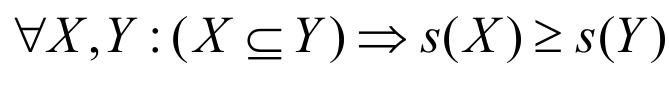


**سوال ششم‌:**

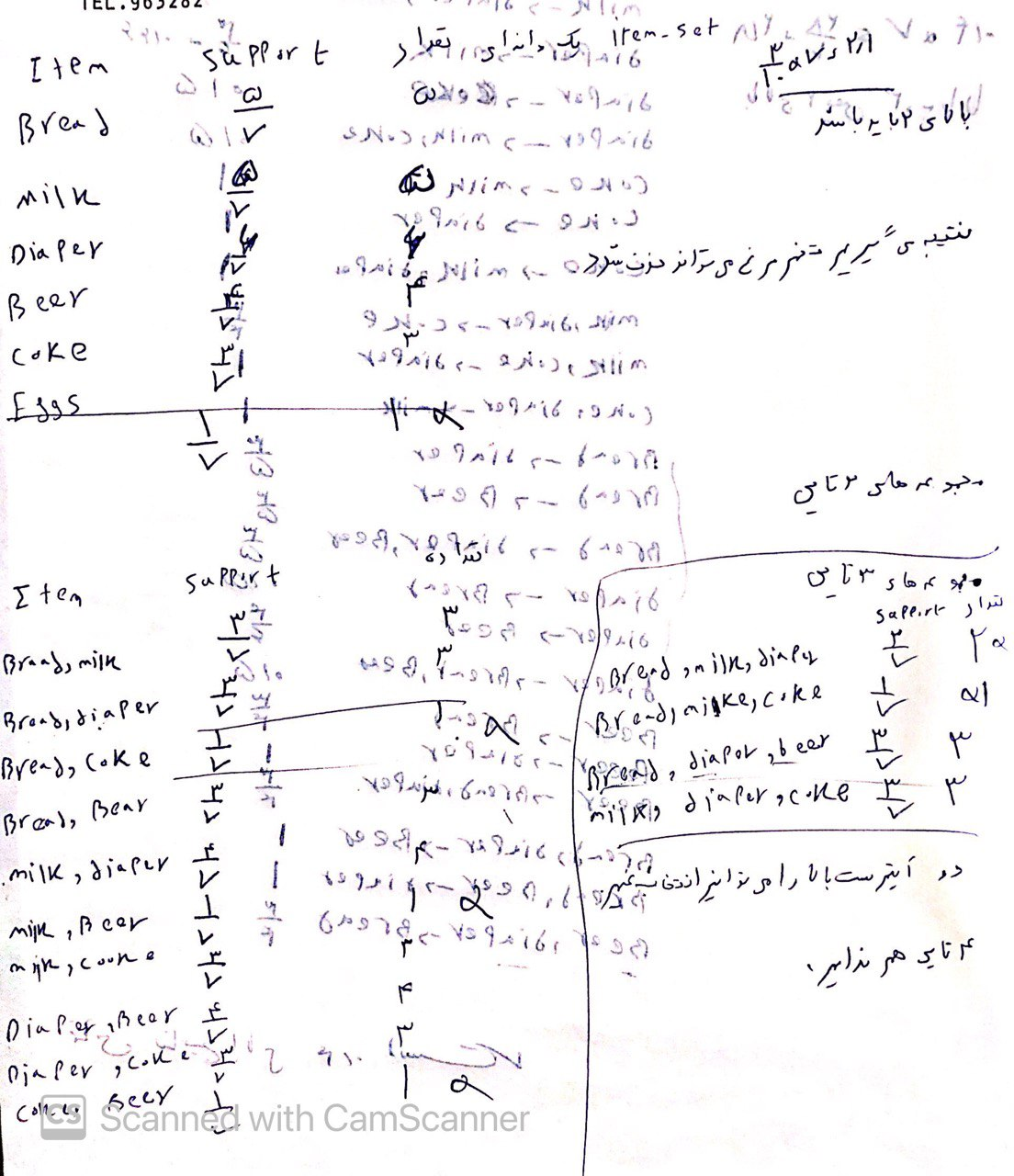


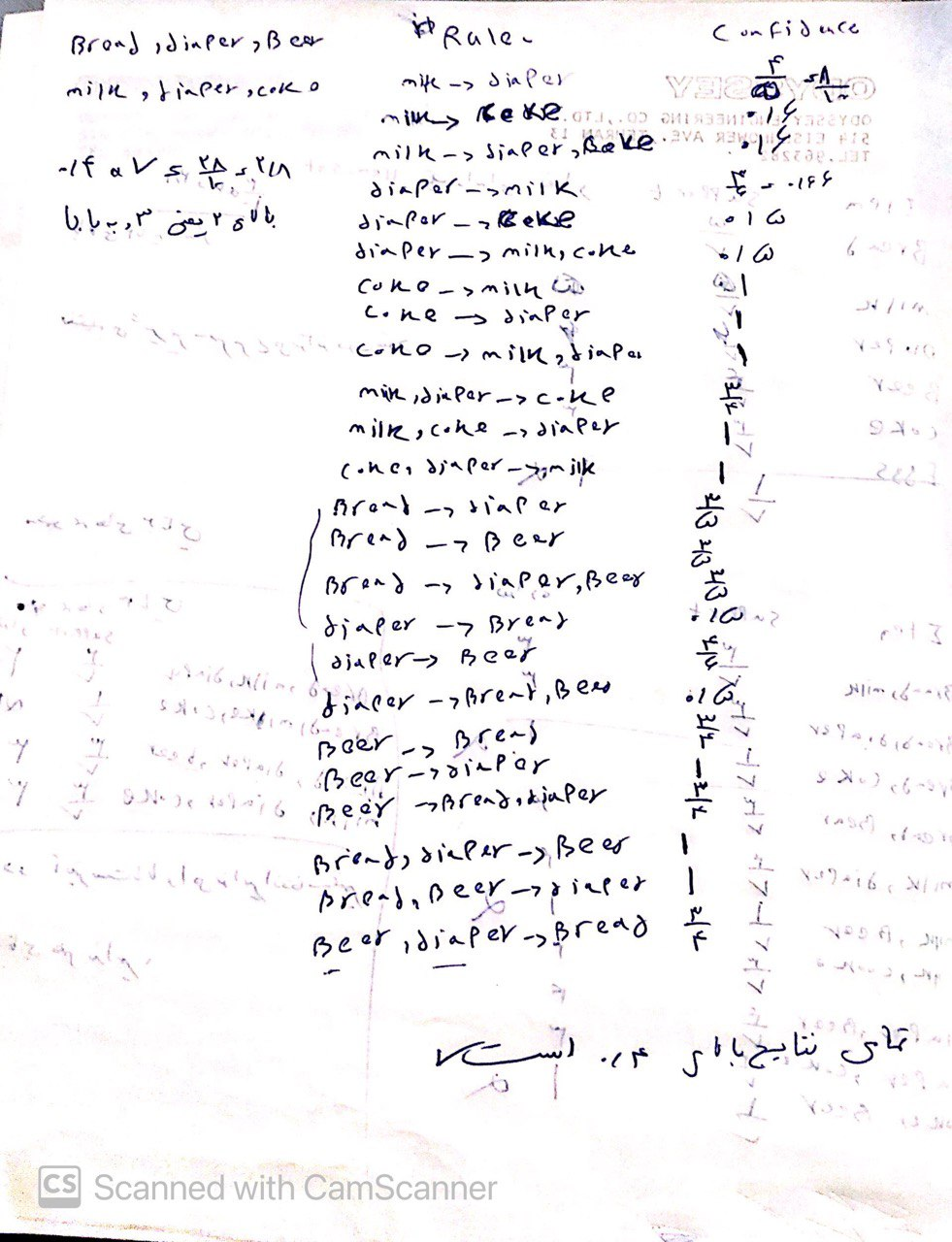
**پاسخ :**

برای اینکه support بالای ۰٫۳ را حساب کنیم میتوانیم از این قانون استفاده کنیم :

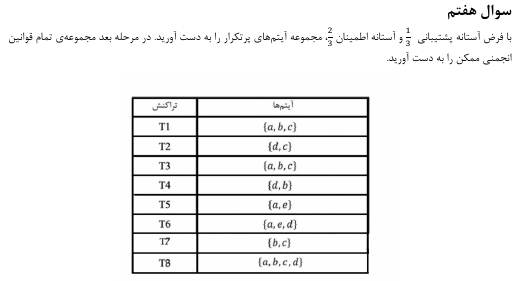


که ابتدا از مجموعه های یک عضوی شروع کنیم و در نهایت آن هایی که بالای ۰٫۳ هستند را حساب کرده و آستانه اطمینان بالا ۰٫۴ را انتخاب کنیم:

ادامه پاسخ در عکس 



**سوال هفتم :**



**پاسخ :**

